# Assignment 3: Data Exploration

Aubrey Knier, Section #2

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on Data Exploration.

## Directions

1. Change "Student Name, Section #" on line 3 (above) with your name and section number.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "FirstLast_A03_DataExploration.Rmd") prior to submission.

The completed exercise is due on <>.

## Set up your R session

1. Check your working directory, load necessary packages (tidyverse), and upload two datasets: the ECOTOX neonicotinoid dataset (ECOTOX_Neonicotinoids_Insects_raw.csv) and the Niwot Ridge NEON dataset for litter and woody debris (NEON_NIWO_Litter_massdata_2018-08_raw.csv). Name these datasets "Neonics" and "Litter", respectively. **Be sure to add the `stringsAsFactors = TRUE` parameter to the function when reading in the CSV files.**

```
getwd()
```

```
## [1] "/Users/aubreyknier/Desktop/Spring 2022/ENV872_EDA/Environmental_Data_Analytics_2022"
```

```
library(tidyverse)

Neonics <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Insects_raw.csv", stringsAsFactors=TRUE)
Litter <- read.csv('./Data/Raw/NEON_NIWO_Litter_massdata_2018-08_raw.csv', stringsAsFactors=TRUE)
```

## Learn about your system

2. The neonicotinoid dataset was collected from the Environmental Protection Agency's ECOTOX Knowledgebase, a database for ecotoxicology research. Neonicotinoids are a class of insecticides used widely in agriculture. The dataset that has been pulled includes all studies published on insects. Why might we be interested in the ecotoxicologoy of neonicotinoids on insects? Feel free to do a brief internet search if you feel you need more background information.

Answer: While neonicotinoids were created to control insects that are harmful to crops, it's important to investigate the effect of these insecticides on non-target insect species. Neonicotinoids are toxic to important pollinators, namely bees, and therefore pose a severe, cascading ecological threat.

3. The Niwot Ridge litter and woody debris dataset was collected from the National Ecological Observatory Network, which collectively includes 81 aquatic and terrestrial sites across 20 ecoclimatic domains. 32 of these sites sample forest litter and woody debris, and we will focus on the Niwot Ridge long-term ecological research (LTER) station in Colorado. Why might we be interested in studying litter and woody debris that falls to the ground in forests? Feel free to do a brief internet search if you feel you need more background information.

   Answer: Falling litter and woody debris are important to forest ecosystems as they eventually decompose and return nutrients to the forest floor, which is required to continue future growth and succession. Studying litter and woody debris may help us track nutrient cycling and forest ecosystem health.

4. How is litter and woody debris sampled as part of the NEON network? Read the NEON_Litterfall_UserGuide.pdf document to learn more. List three pieces of salient information about the sampling methods here:

   Answer: 1. Sampling occurred exclusively at tower plots in terrestrial NEON sites with woody vegetation >2m tall.
   2. The placement of litter traps was random in sites with over 50% aerial cover of woody vegetation >2m in height, and targeted in sites with less than 50% cover of woody vegetation.
   3. Ground traps were sampled once a year and elevated traps were sampled once every 2 weeks in deciduous forest sites during senescence and once every 1-2 months at evergreen sites.

## Obtain basic summaries of your data (Neonics)

5. What are the dimensions of the dataset?

```
dim(Neonics)
```

```
## [1] 4623   30
```

6. Using the `summary` function on the "Effect" column, determine the most common effects that are studied. Why might these effects specifically be of interest?

```
summary(Neonics$Effect)
```

```
##     Accumulation        Avoidance         Behavior     Biochemistry
##               12              102              360               11
##          Cell(s)      Development        Enzyme(s) Feeding behavior
##                9              136               62              255
##         Genetics           Growth        Histology       Hormone(s)
##               82               38                5                1
##    Immunological     Intoxication       Morphology        Mortality
##               16               12               22             1493
##       Physiology       Population     Reproduction
##                7             1803              197
```

Answer: The most common effects studied were Mortality and Population. These are likely of specific interest to determine which insect species neonicotinoids are killing and how the insecticides affect species abundance.

7. Using the `summary` function, determine the six most commonly studied species in the dataset (common name). What do these species have in common, and why might they be of interest over other insects? Feel free to do a brief internet search for more information if needed.

```
summary(Neonics$Species.Common.Name)
```

```
##                          Honey Bee                      Parasitic Wasp
##                                667                                285
##               Buff Tailed Bumblebee                Carniolan Honey Bee
##                                183                                152
##                         Bumble Bee                   Italian Honeybee
##                                140                                113
##                    Japanese Beetle                  Asian Lady Beetle
##                                 94                                 76
##                     Euonymus Scale                           Wireworm
##                                 75                                 69
##                  European Dark Bee                  Minute Pirate Bug
##                                 66                                 62
##                Asian Citrus Psyllid                     Parastic Wasp
##                                 60                                 58
##              Colorado Potato Beetle                   Parasitoid Wasp
##                                 57                                 51
##                 Erythrina Gall Wasp                      Beetle Order
##                                 49                                 47
##         Snout Beetle Family, Weevil           Sevenspotted Lady Beetle
##                                 47                                 46
##                     True Bug Order                Buff-tailed Bumblebee
##                                 45                                 39
##                       Aphid Family                    Cabbage Looper
##                                 38                                 38
##                Sweetpotato Whitefly                     Braconid Wasp
##                                 37                                 33
##                      Cotton Aphid                     Predatory Mite
##                                 33                                 33
##              Ladybird Beetle Family                        Parasitoid
##                                 30                                 30
##                      Scarab Beetle                     Spring Tiphia
##                                 29                                 29
##                       Thrip Order                Ground Beetle Family
##                                 29                                 27
##                  Rove Beetle Family                     Tobacco Aphid
##                                 27                                 27
##                      Chalcid Wasp               Convergent Lady Beetle
##                                 25                                 25
##                     Stingless Bee                  Spider/Mite Class
##                                 25                                 24
##                 Tobacco Flea Beetle                   Citrus Leafminer
##                                 24                                 23
##                    Ladybird Beetle                         Mason Bee
```

```
##                              23                              22
##                        Mosquito                 Argentine Ant
##                              22                              21
##                          Beetle      Flatheaded Appletree Borer
##                              21                              20
##             Horned Oak Gall Wasp             Leaf Beetle Family
##                              20                              20
##                Potato Leafhopper      Tooth-necked Fungus Beetle
##                              20                              20
##                     Codling Moth       Black-spotted Lady Beetle
##                              19                              18
##                     Calico Scale            Fairyfly Parasitoid
##                              18                              18
##                      Lady Beetle          Minute Parasitic Wasps
##                              18                              18
##                        Mirid Bug                Mulberry Pyralid
##                              18                              18
##                         Silkworm                  Vedalia Beetle
##                              18                              18
##            Araneoid Spider Order                       Bee Order
##                              17                              17
##                   Egg Parasitoid                   Insect Class
##                              17                              17
##          Moth And Butterfly Order     Oystershell Scale Parasitoid
##                              17                              17
## Hemlock Woolly Adelgid Lady Beetle           Hemlock Wooly Adelgid
##                              16                              16
##                            Mite                    Onion Thrip
##                              16                              16
##             Western Flower Thrips                   Corn Earworm
##                              15                              14
##                Green Peach Aphid                      House Fly
##                              14                              14
##                         Ox Beetle              Red Scale Parasite
##                              14                              14
##                Spined Soldier Bug           Armoured Scale Family
##                              14                              13
##                 Diamondback Moth                  Eulophid Wasp
##                              13                              13
##                Monarch Butterfly                  Predatory Bug
##                              13                              13
##             Yellow Fever Mosquito            Braconid Parasitoid
##                              13                              12
##                     Common Thrip     Eastern Subterranean Termite
##                              12                              12
##                           Jassid                     Mite Order
##                              12                              12
##                         Pea Aphid               Pond Wolf Spider
##                              12                              12
##          Spotless Ladybird Beetle          Glasshouse Potato Wasp
##                              11                              10
##                         Lacewing          Southern House Mosquito
##                              10                              10
##           Two Spotted Lady Beetle                     Ant Family
```

```
##                         10                          9
##               Apple Maggot                    (Other)
##                          9                        670
```

Answer: The six most commonly studied species in the dataset are Honey Bee, Parasitic Wasp, Buff Tailed Bumblebee, Carniolan Honey Bee, Bumble Bee, Italian Honeybee. These species are almost all bees, which are of special interest because a decline in pollinators can significantly impact global ecology and our food supply.

8. Concentrations are always a numeric value. What is the class of Conc.1..Author. in the dataset, and why is it not numeric?

```
class(Neonics$Conc.1..Author.)
```

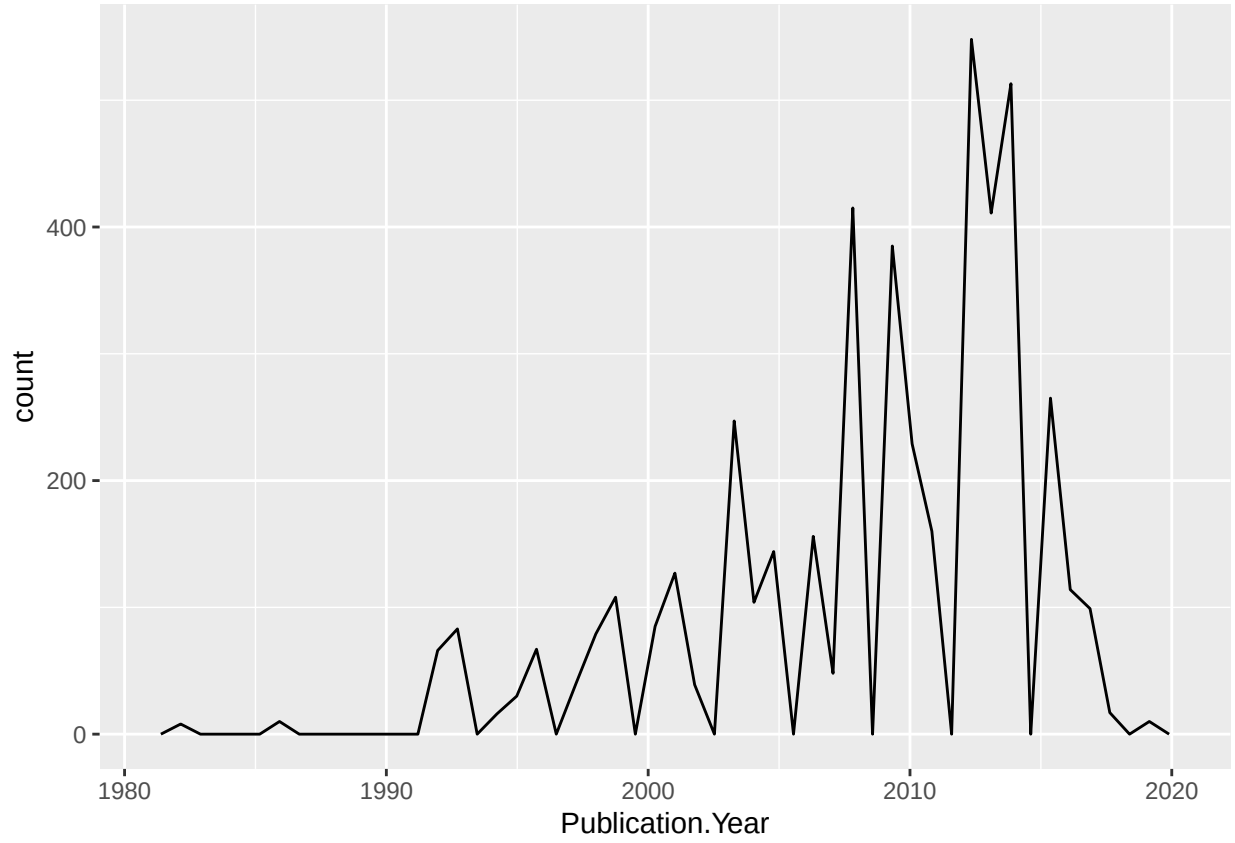
```
## [1] "factor"
```

Answer: The class of "Conc.1..Author" is a factor within the dataset. It is not numeric because there are "NR"s and special characters listed.

## Explore your data graphically (Neonics)

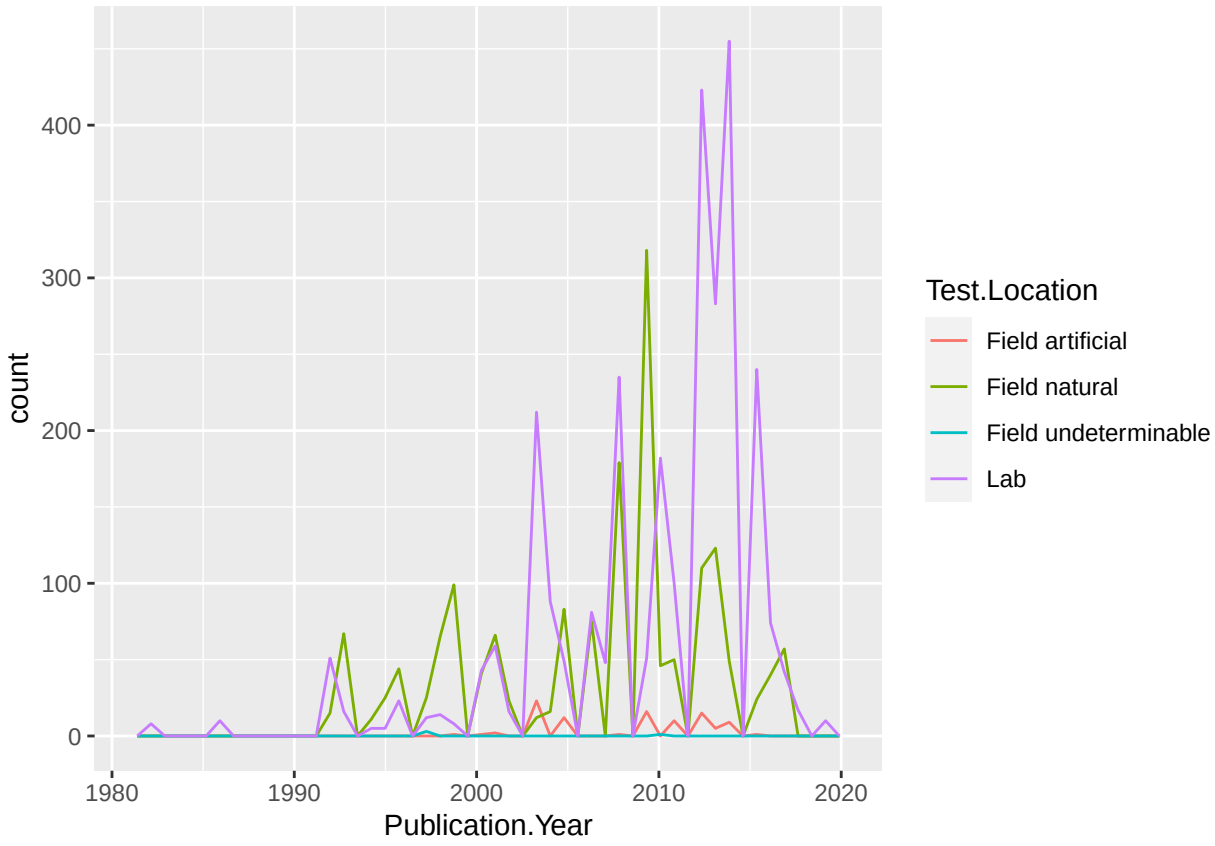
9. Using `geom_freqpoly`, generate a plot of the number of studies conducted by publication year.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year), bins = 50)
```



10. Reproduce the same graph but now add a color aesthetic so that different Test.Location are displayed as different colors.

```
ggplot(Neonics) +
  geom_freqpoly(aes(x = Publication.Year, color=Test.Location), bins = 50)
```

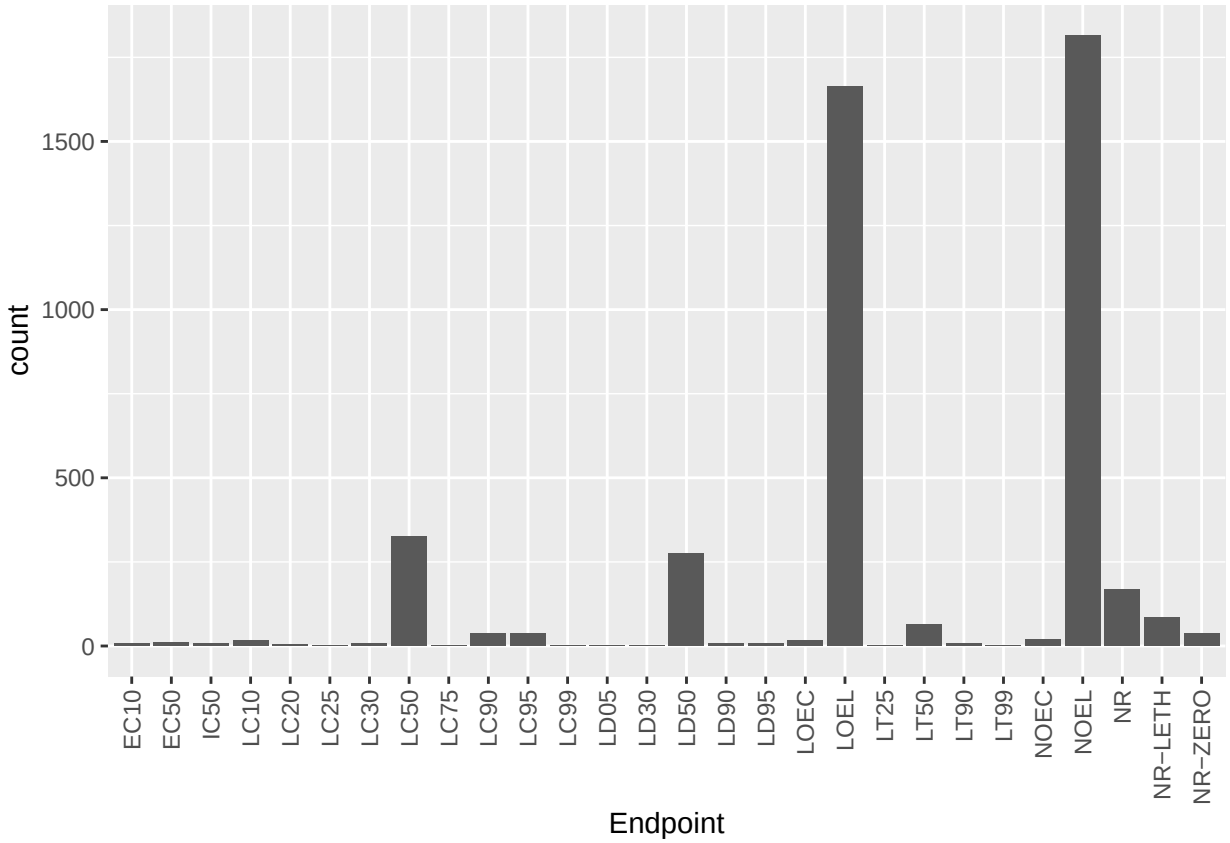



Interpret this graph. What are the most common test locations, and do they differ over time?

> Answer: The most common test locations are "Lab" and "Field natural" and the number of studies conducted in each location differ over time. For example, around 2009, "Field natural" was most common, but after 2010, "Lab" became the most popular test location by far.

11. Create a bar graph of Endpoint counts. What are the two most common end points, and how are they defined? Consult the ECOTOX_CodeAppendix for more information.

```
ggplot(Neonics, aes(x = Endpoint)) +
  geom_bar() +
    theme(axis.text.x=element_text(angle=90, vjust=0.5, hjust=1))
```




Answer: The two most common endpoints are NOEL and LOEL. NOEL, or "no-observable-effect-level", means that the highest dose produced effects that were not significantly different from responses of controls. LOEL, or "lowest-observalbe-effect-level", means that the lowest dose produced effects that were significantly different from responses of controls.

## Explore your data (Litter)

12. Determine the class of collectDate. Is it a date? If not, change to a date and confirm the new class of the variable. Using the `unique` function, determine which dates litter was sampled in August 2018.

```
class(Litter$collectDate)
```

```
## [1] "factor"
```

```
Litter$collectDate <- as.Date(Litter$collectDate)
unique(Litter$collectDate)
```

```
## [1] "2018-08-02" "2018-08-30"
```

13. Using the `unique` function, determine how many plots were sampled at Niwot Ridge. How is the information obtained from `unique` different from that obtained from `summary`?

```r
unique(Litter$plotID) #names of plots samples
```

```
##  [1] NIWO_061 NIWO_064 NIWO_067 NIWO_040 NIWO_041 NIWO_063 NIWO_047 NIWO_051
##  [9] NIWO_058 NIWO_046 NIWO_062 NIWO_057
## 12 Levels: NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 ... NIWO_067
```

```r
length(unique(Litter$plotID)) #how many plots sampled
```

```
## [1] 12
```

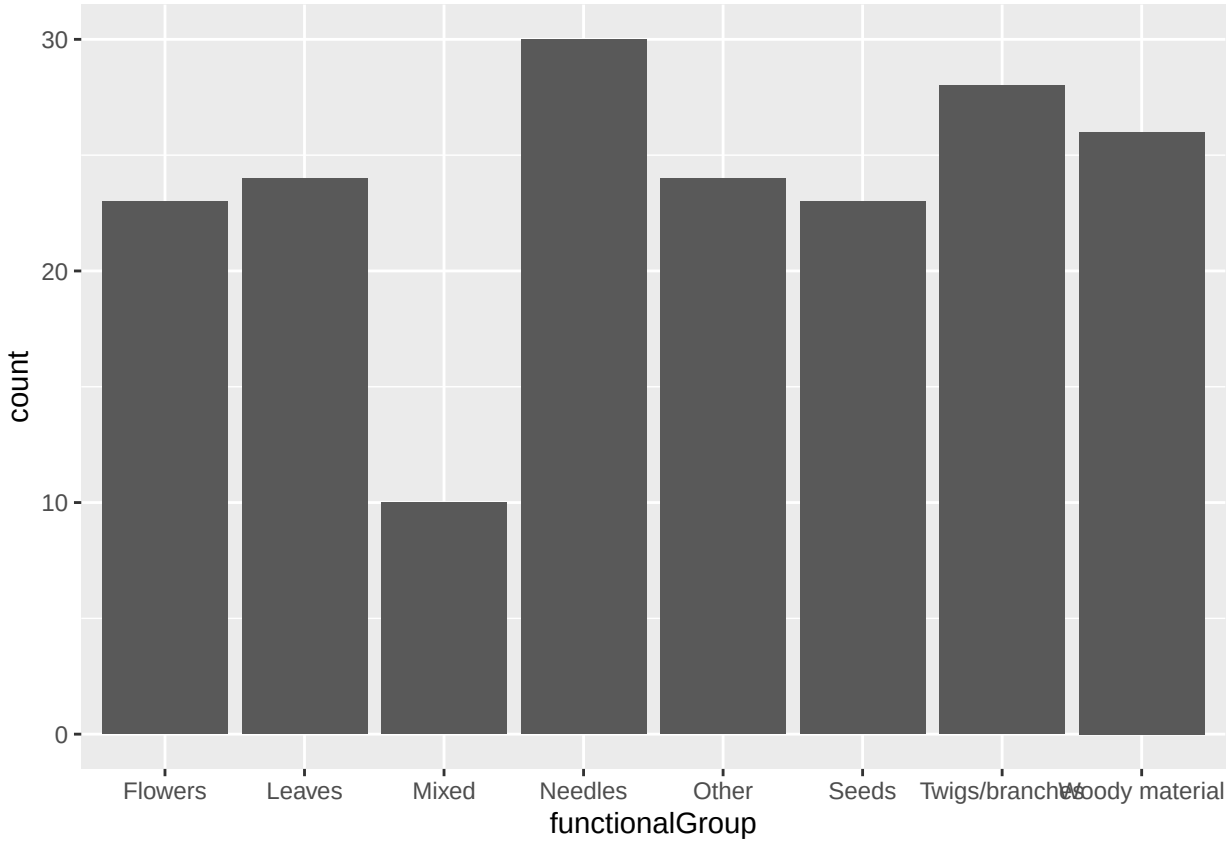
```r
summary(Litter$plotID) #testing result of summary function
```

```
## NIWO_040 NIWO_041 NIWO_046 NIWO_047 NIWO_051 NIWO_057 NIWO_058 NIWO_061
##       20       19       18       15       14        8       16       17
## NIWO_062 NIWO_063 NIWO_064 NIWO_067
##       14       14       16       17
```

Answer: The unique function returns a list of each element that occurs at least once in the column. Therefore, unique() returns a simple list of the plot names. The summary function, however, provides a count of how many samples there are from each plot, which means that summary() also provides a list of plot names, but with added information.

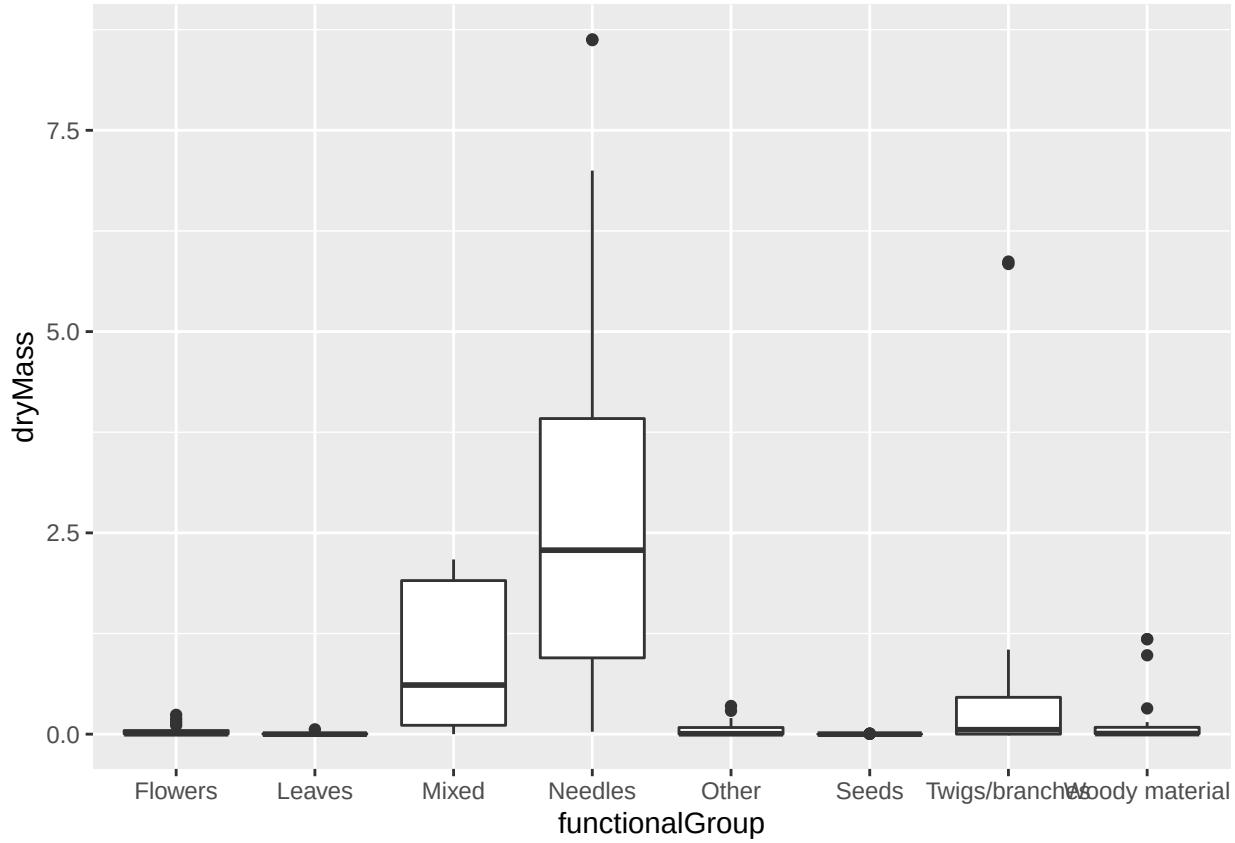
14. Create a bar graph of functionalGroup counts. This shows you what type of litter is collected at the Niwot Ridge sites. Notice that litter types are fairly equally distributed across the Niwot Ridge sites.

```r
ggplot(Litter, aes(x = functionalGroup)) +
  geom_bar()
```




15. Using `geom_boxplot` and `geom_violin`, create a boxplot and a violin plot of dryMass by functionalGroup.

```
#boxplot
ggplot(Litter) +
  geom_boxplot(aes(x = functionalGroup, y = dryMass))
```

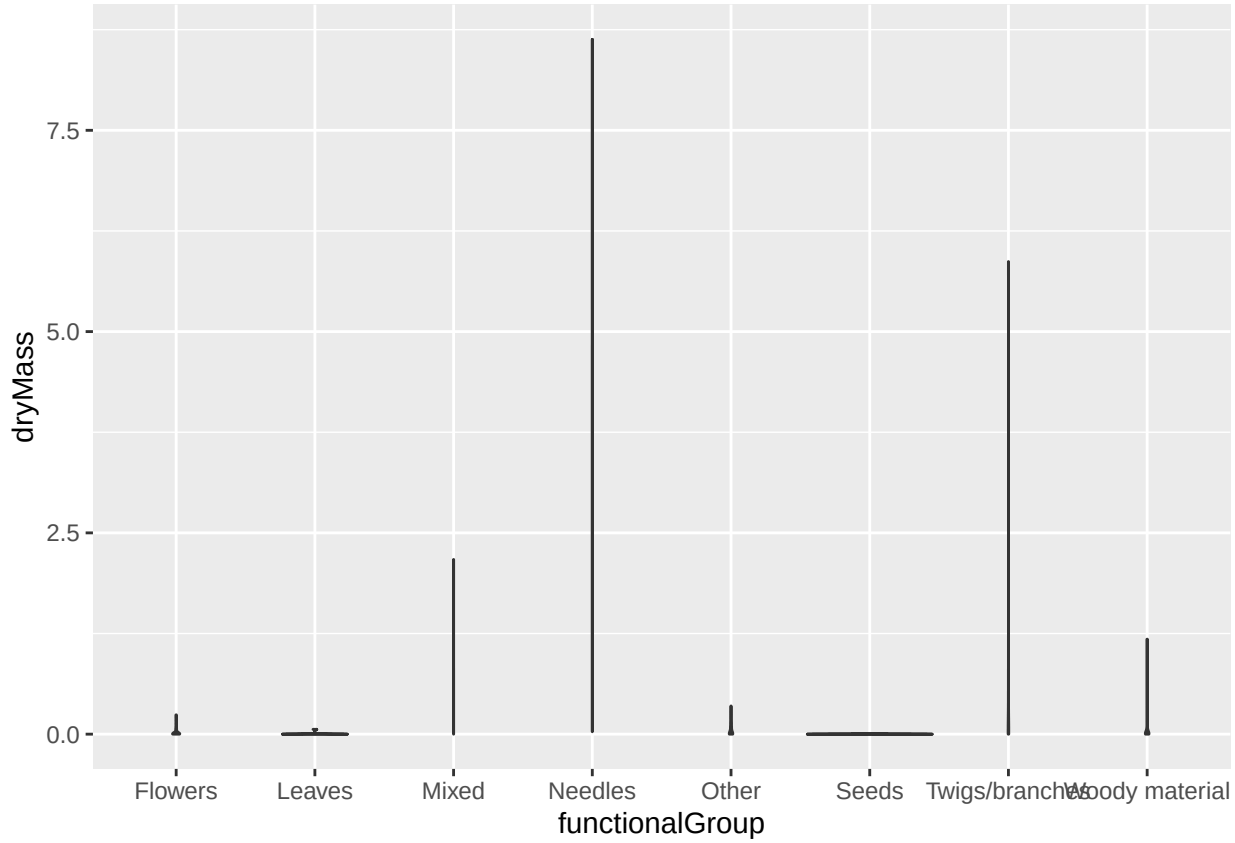



```
#violin plot
ggplot(Litter) +
  geom_violin(aes(x = functionalGroup, y = dryMass),
     draw_quantiles = c(0.25, 0.5, 0.75)) #shows IQR and width indicates how many data points are in it
```

```
## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values

## Warning in regularize.values(x, y, ties, missing(ties), na.rm = na.rm):
## collapsing to unique 'x' values
```




```
#finding median biomass
median(Litter$dryMass[Litter$functionalGroup=="Needles"])
```

```
## [1] 2.285
```

```
median(Litter$dryMass[Litter$functionalGroup=="Mixed"])
```

```
## [1] 0.61
```

Why is the boxplot a more effective visualization option than the violin plot in this case?

> Answer: The violin plot is not a helpful visualization for dry mass by litter types because for groups such as "Needles", "Mixed", and "Twigs/branches", the density of points is so low throughout the plot that it appears as simply a vertical line. Thus, a boxplot is preferrable because it displays Q1, Median, and Q3 regardless of distribution, so more information is visibly displayed for these litter types. Also, it's important to note that the ranges are extremely variable among groups, which makes groups with smaller ranges very difficult to read in both types of plots. However, they are slightly easier to visualize in the boxplots.

What type(s) of litter tend to have the highest biomass at these sites?

> Answer: Needle litter tends to have the highest biomass, with a median dry mass value of 2.29. Mixed litter has the second highest biomass, with a median dry mass value of 0.61.